Robust Methods, Visual Displays, & Statistical Computing: Critical Statistical Methods for Big Data

Karen Kafadar

Department of Statistics, UVA

Statistical Computing and Robust Inference for High-Dimensional Data: Taipei, Taiwan

kkafadar@virginia.edu

http://statistics.as.virginia.edu/faculty-staff/profile/kk3ab

With thanks to Jordan Rodu, UVA

and to IASC, founded 1977

Acknowledgements: NIJ, CSAFE (NIST, via Iowa State)

Motivation for this talk: Victims of Sexual Abuse

Prof Kathryn Laughon, School of Nursing, UVA

- Victims of abuse require thorough forensic examination
- Sexual Assault only (SA): Attacker free on his own recognizance
- Evidence of (attempted) Strangulation (ST), with or w/o SA? Much more serious: attempted murder charge, usually jail time
- Courtroom: “What leads you to believe ST was attempted?”

- Laughon (forensic nursing): “Well, *in my experience, ...* ”
- Judge: “Mere ‘impressions’ are inadequate; Verdict = SA”
- Attacker not charged with ST. What does he do?
- *What “proof” is needed to establish when ST was attempted?*

Need data to quantify features distinguishing ST from SA.

Simple classification problem?

- Arizona database of 12,099 victims
- 5,784 SA-only; 5,469 ST-only, 846 ST+SA
- ST, with or without SA, is a felony (jail time); hence, combine (relatively few) ST+SA cases with ST-only
- Lots of data: 46% white, 29% Hisp, 25% other; $\sim$30 injury types, $\sim$45 injury locations (disassociated)
- Combined locations (e.g., ears + eyes + nose + face + mouth)

Single-feature screening: Not a single one of these variables showed significantly different proportions in the 2 classes.
(??? Makes no sense: expect 3-4 by chance alone!)

OUTLINE

1. Are big data *informative* or *deceiving*? [either]
2. With big data, do we need robust methods & statistical displays? [yes, now more than ever]
3. When do we need *efficient* statistical computing?
4. How much effort into black-box (AI/ML) algorithms should statisticians devote? When / When NOT to use them?
5. Statistics Now & Later

1. Big data: informative or deceiving?

Statistical problems of the Kinsey report, 1953
Cochran, Mosteller, Tukey: *JASA* 48: 673-716

- Rockfeller Foundation (RF) funded Indiana Univ researcher Alfred Kinsey's studies *Sexual Behavior in the Human Male*
- Unexpected results were widely publicized
- RF, via NRC, asked Cochran, Mosteller, Tukey to examine "statistical and methodological issues"
- Report was balanced, in reporting aspects of the study that were done well — *but also many others that were not*

Among the problems:

- Self-selected study participants
- Biased, inconsistent data: Oral questionnaire, likely varied "*form of questions to suit the subject and the situation*"
- No consideration of obtaining probability sample
- "Statistical analysis" of data *as if* they came from a *random* (!) sample
- Accuracy of results *impossible* to quantify (surely not $1/\sqrt{n}$)
- No accounting for, or understanding causes of, refusers

Have these problems multiplied with current data?

Kinsey Report proposed recommendations:

- Recognize differences among *reported, recorded*, and *observed* behavior. Only *reported* behavior is available ⇒ analysis has systematic bias in *measurement*;
- *recorded* or *observed* data likely have *systematic errors in sampling* – which may exceed errors in *measurement*;
- Possible approaches to estimating quantities from non-probability samples;
- Other metrics for measuring behavior of interest;
- More data checks;
- Adjustment to reduce bias;
- Methods for obtaining more "probability-like" samples

Principles of Sampling: Cochran, Mosteller, Tukey *JASA* 1954

NYT Obituary of J.W. Tukey, 28 July 2000:

“*In a series of meetings over two years, Mr. Kinsey vigorously defended his work, which Mr. Tukey believed was seriously flawed, relying on a sample of people who knew each other. Mr. Tukey said a random selection of three people would have been better than a group of 300 chosen by Mr. Kinsey.*”

Fast forward to 2018: Xiao-Li Meng, *AOAS* 2018:

X-L Meng, “Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 U.S. Presidential election,” *AOAS* 2018:12(2),685-726

“*Almost since the dawn of statistics, the dominating mathematical tool for justifying statistical methods has been large-sample asymptotics. Neither the Law of Large Numbers nor the Central Limit Theorem, two pillars of the vast statistical palace, could be established without such asymptotics. Surely then we statisticians must be thrilled by the explosive growth of data size ...*
A statistical paradise would seem to have arrived.

“*The reality appears to be the opposite.*”

Which is more trustworthy, a 5% survey sample or 80% sampling of [tons of] administrative records?

N = population size; n = sample size;
$X_j = j^{th}$ response, $j \in I_n \subset \{1,...,N\}$; $R_j = \mathrm{P}\{X_j$ is observed$\}$ (binary)
$\rho_{R,G} \equiv \mathrm{corr}(R_j, G(X_j))$, $G(\cdot)$ = desired function of X

Target: $\bar{G}_N$ = Average of $G(X)$ for all $X = 1, ..., N$.

- Estimate $\bar{G}_N$ by $\bar{G}_n = \sum_{j=1}^{N} R_j X_j \sum_{j=1}^{N} R_j$
- Joint distribution of $\mathbf{R} = \{R_1, ..., R_N\}$:
  - Well-specified for *probabilistic random sampling*
  - No probabilistic mechanism for *recorded* or *self-reported* data
- We generally assume that $\rho_{R,G} = 0$. $\mathrm{P}\{X_j$ is observed $\mid R_j\} = 0$.
  What if $\rho_{R,G} \neq 0$?

To assess bias & MSE of $\bar{G}_n$, need to incorporate $\mathbf{R}$.

$$\text{bias} = \bar{G}_n - \bar{G}_N = Cov(R,G)/E(R) = \rho_{R,G} \times \sqrt{(1-f)/f} \times \sigma_G$$
$$\Rightarrow (\bar{G}_n - \bar{G}_N)/\sqrt{Var_{SRS}(\bar{G}_n)} = \sqrt{N-1}\rho_{R,G}$$

where $E(\cdot), Cov(\cdot,\cdot)$ is wrt Uniform distribution on $\{1,...,N\}$;
$E_{\mathbf{R}}$ wrt $\mathbf{R}$; $f = fpc = [(N-n)/(N-1)]^{1/2}$ (often close to 1).

For SRS: $E_{\mathbf{R}}(\rho^2_{R,G}) = 1/(N-1)$

In general, $\propto 1/N$ *for probability sampling only.*

Else, *Relative error* is $\propto \rho \times N$.

n = 1M and N = 100M: Tiny $\rho = 0.005 \Rightarrow$ *Relative error is huge*

"*Estimates obtained from the Cooperative Congressional Election Study (CCES) of the 2016 US presidential election suggest a* $\rho_{R,X} \approx -0.005$ *for self-reporting to vote for Donald Trump. … This seemingly minuscule data defect correlation implies that the simple sample proportion of the self-reported voting preference for Trump from 1% of the US eligible voters [n = 2.3M] has the same mean squared error as the corresponding sample proportion from a genuine simple random sample of size* $n \approx 400$*, a 99.98% reduction of sample size (and hence our confidence).*"

"Big" data may be very misleading.

Sexual assault study: Which SA cases involve Strangulation (ST)?

- AZ database of 12,099 cases; drop 3644 (very incomplete)
- 4,329 (SA-only; 45.8%) + 4,418 (ST-only: 46.7%) + 708 (ST+SA: 7.5%) = 9,455
- 11 injury types × 4 locations (upper/lower body, head, neck)
- Forensic nurse: "*petechiae highly indicative of ST*"
- Data: petechiae present in nearly equal proportions in ST/SA: 13.5%/16.1% (head); 17.8%/20.8% (neck)
- Nothing even close to significant?

Suspect "scrambling" of rows, impossible to disentangle

Advice: “Conduct forensic exams using consistent criteria and simple recording forms directly into computer”

- All 170 exams at UVA Health under guidance of Dr. Laughon
- 77 (SA-only: 45.3%), 74 (ST-only: 43.5%), 19 (ST+SA: 11.2%)
- **77 SA-only** vs **93 ST/ST+SA** (“*ST*”)
- Among SA-only: No petechiae in 76 (98.7%); single occurrence in 1
- Among ST: No petechiae in 70 (75.2%); single occurrence in 11 (11.8%), 2+ occurrences on 12 (12.9%)
- ST cases had more injuries than SA-only cases:
  16 SA-only (20.7%) vs 63 ST (67.7%) had 5+ injuries overall
- These results made more sense.

Note: *When injuries exist in SA-only, usually only 1 or 2*:

Head+Neck:

	#injuries				
	0	1	2	3+	(% > 0)
SA	65	7	3	2	(16%)
ST	22	17	14	40	(76%)

Face (ears, eyes, nose):

	#injuries				
	0	1	2	3+	(% > 0)
SA	67	6	2	2	(13%)
ST	41	17	8	27	(56%)

Swelling:

	#injuries				
	0	1	2	3+	(% > 0)
SA	73	4	0	0	(5%)
ST	64	21	0	0	(31%)

Petechiae: Very informative

	#injuries				
	0	1	2	3+	(% > 0)
SA	76	1	0	0	(1%)
ST	70	11	4	8	(25%)

“Machine learning” approach?

Tree (`rpart` in `R`) indicates only two variables:

Head+neck (0 or > 0), Face (0 or > 0)

- Head+Neck > 0: “ST” (71 correct, 12 incorrect)
- Head+Neck $= 0$ & Face > 0: “ST” (10 correct, 7 incorrect)
- Head+Neck $= 0$ & Face $= 0$: “SA” (58 correct, 12 incorrect)

	Algorithm		
True	“SA”	“ST”	%Error
SA	58	19	(25%)
ST	12	81	(13%)

FPR: 1 in 4 SA cases falsely accused of ST (too high)

Statistical analysis: Logistic regression;
Sensible, biology-driven combinations of locations

Assess a case as "ST" if:

1. *Petechiae is present* or

2. *Petechiae is absent but injuries to the mouth, face, head, neck total* **L** *or more,* **or** *the number is less than* **L** *and the number of* **all** *injuries exceeds 14.*

When $\mathbf{L = 3}$, simple algorithm correctly predicts, on average:
68% of 93 ST/ST+SA cases (63 correct, 30 incorrect): FNR = 32.0%
95% of 77 SA-only cases (73 correct, 4 incorrect): FPR = 5.2%.

"**LR**" = 5.8 (FPR, FNR confirmed by bootstrap & cross-validation)

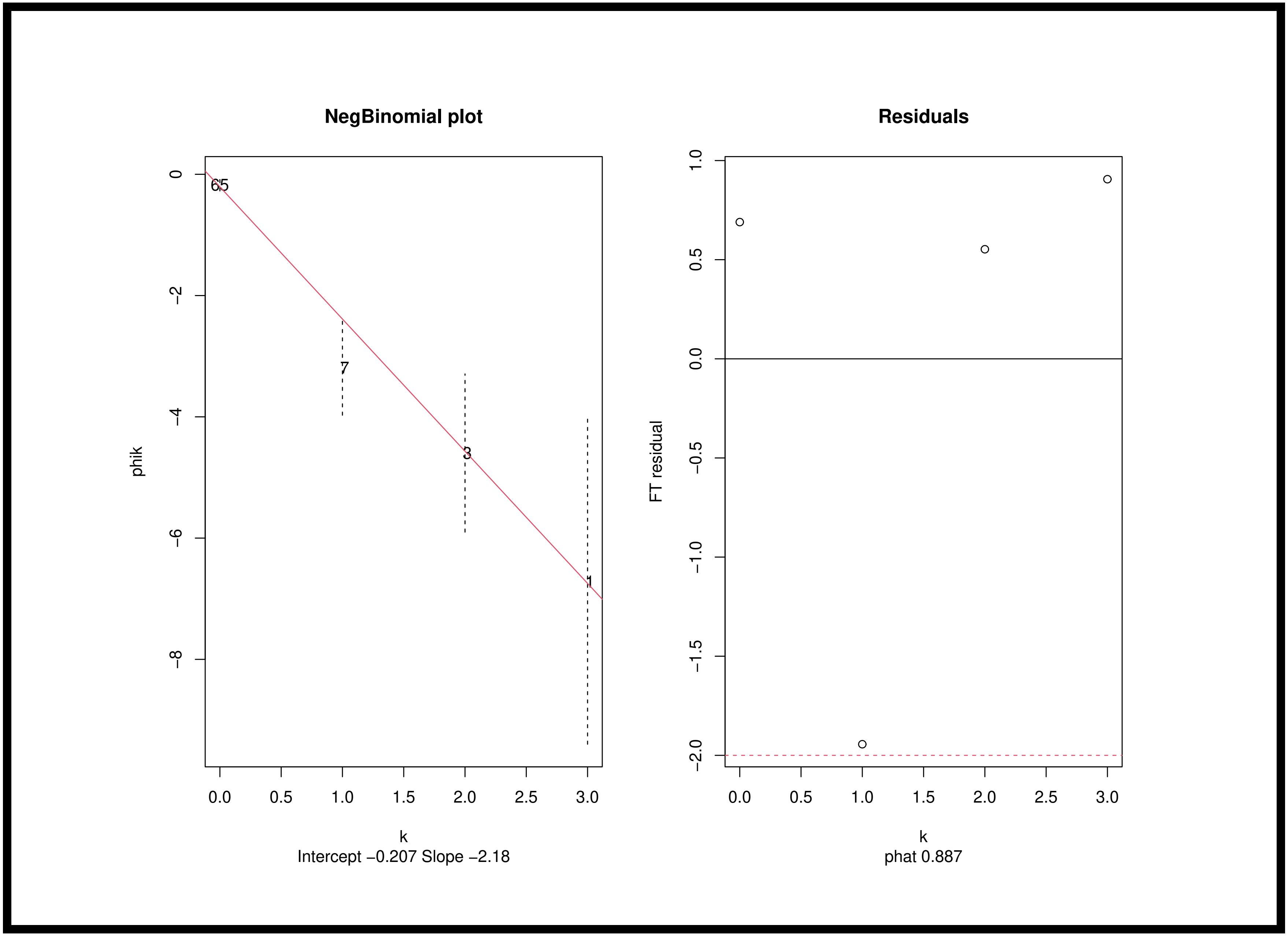
Fitting distributions of #injuries in ST/ST cases:

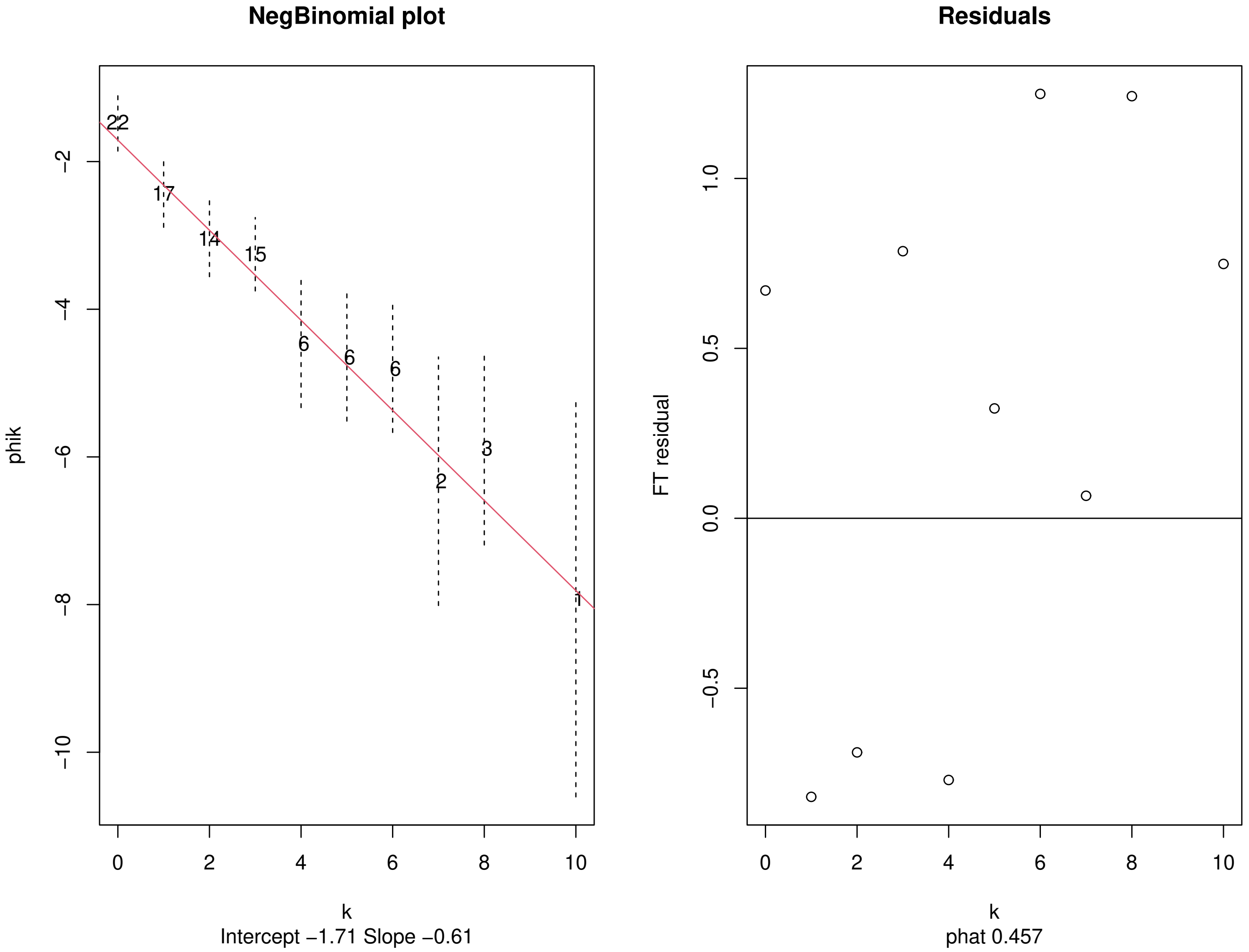
DC Hoaglin, Ch.7 in *Exploring Data Tables, Trends, and Shapes*

Robust method for fitting Poisson, Binomial, Negative Binomial,...

Negative Binomial(n, p) provided good fit #injuries;
where (n, p) differs for ST vs SA, likely to be a good predictor

	SA	ST
Head+Neck	NB($n = 2$, $\hat{p} = 0.887$)	NB($n = 2$, $\hat{p} = 0.457$)
Upper Body	NB($n = \frac{1}{4}$, $\hat{p} = 0.172$)	NB($n = 1$, $\hat{p} = 0.262$)
Total Injuries	NB($n = 1$, $\hat{p} = 0.257$)	NB($n = 1$, $\hat{p} = 0.136$)


NegBinomial plot
phik
0
−2
−4
−6
−8
65
7
3
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
k
Intercept −0.207 Slope −2.18
Residuals
FT residual
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
k
phat 0.887


NegBinomial plot
phik
−2
−4
−6
−8
−10
22
17
14
15
6
6
6
2
3
1
0
2
4
6
8
10
k
Intercept −1.71 Slope −0.61
Residuals
FT residual
1.0
0.5
0.0
−0.5
0
2
4
6
8
10
k
phat 0.457

For this analysis, we relied on:

- Robust methods: Errors in counts of injuries or locations; possibly misclassified cases (ST $\leftrightarrow$ SA)
- Efficient computing (esp with AZ database)
- Reducing high-dimensional data ($n = 170$, $p = 73$) to fewer variables (lower dimensions)
- Sensible biology and medical expertise

Explain algorithm and study results to Judge? (Visual displays)

Penalized Linear Regression Estimators

$$\hat{\beta}_L(\lambda) = \text{argmin Loss}(\beta; y, X) + \lambda \cdot \text{ Penalty}(\beta)$$
$$\hat{\beta}_L(t) = \text{argmin } L(\beta; y, X) \text{ subject to } P(\beta) \leq t$$

Ensure only *some* βs are 'active'

- ST/SA case: We *expect* a specific $\beta_j > 0$
- Another case: GvHD in stem cell transplants: Dr identified
  (A) "I'd be surprised if these variables weren't in model"
  (B) "I'd be surprised if these variables were active"
  (C) "I'd not be surprised either way"
- Enforce $P_A(\beta) \geq t_A$, $P_B(\beta) \leq t_B$, $P_C(\beta) \leq t_C$ $(\geq t_B)$?

Naïvely: Fit A-variables first ⇒ residuals ⇒ penalized regression

2. Robust methods, Statistical displays?

Robust Methods

- Recall: “Robust” to *either*:
  - a *small* fraction of the data are grossly contaminated
    ‘One-Wild’ dist’n: $(n-1)$ from N(0,1), 1 from N(0,100)
  - a *large* fraction of the data slightly contaminated
    e.g., rounded, uncalibrated, (large) interval-data
  - Possibly records from populations of no interest
    (e.g., non-ST/SA cases; external cases among internal ones)
- Robust methods force us to consider: is sample representative of population of interest, from multiple populations?

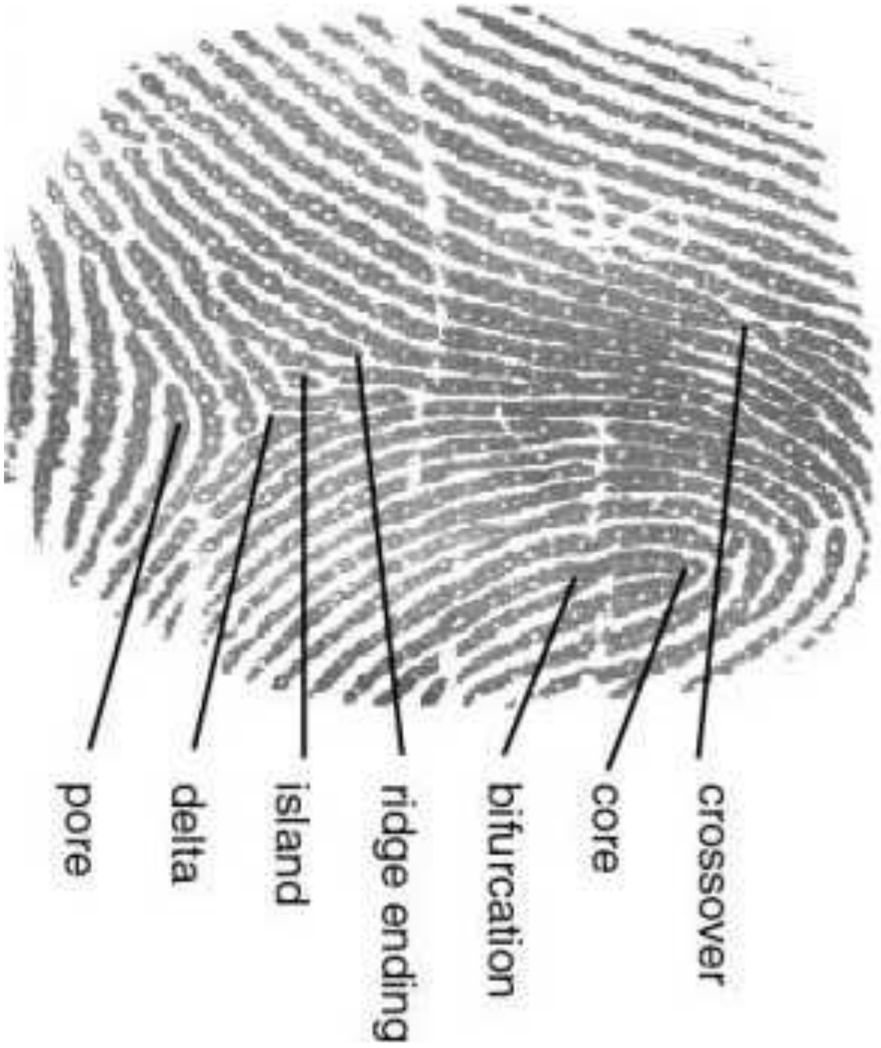
Example from Forensic Science that demonstrates need for, and importance of, sampling high-dimensional data:

Testing algorithms for identifying source of latent print

The task:

- Develop ID algorithm: Which print in a database (NGI: millions of prints) is the source?
- Or: Develop algorithm to measure 'quality' of evidence (poor print ⇒ higher probability of error)
- Very high dimensional (fingerprint features)

Testbed of latent print images: NIST Database 302:
Images of fingerprints from many volunteers



pore
delta
island
ridge ending
bifurcation
core
crossover

Can we use NIST 302 to test an algorithm?

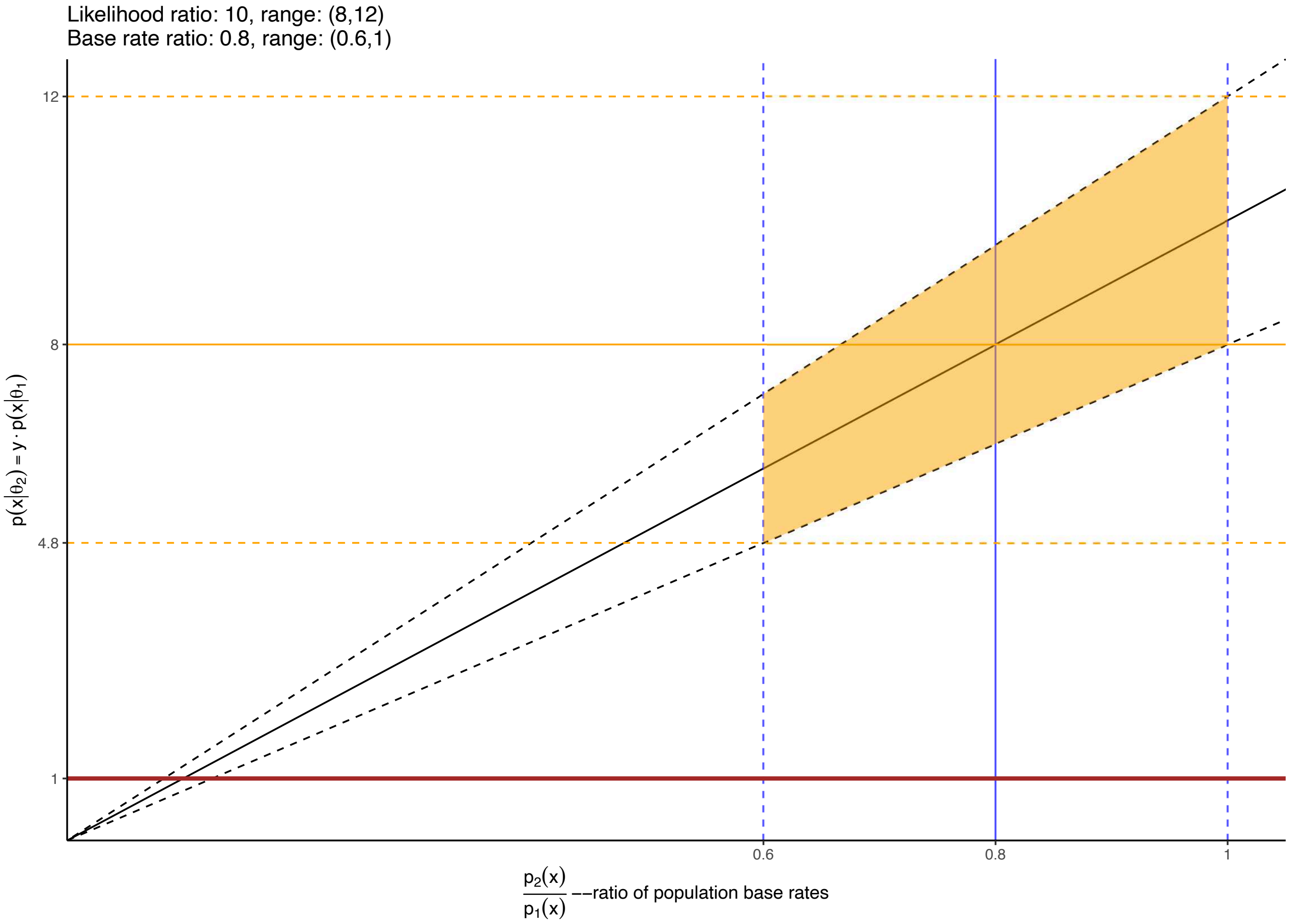
- Is NIST 302 “representative” of *all* fingerprints? (NGI)
- Perhaps, if frequency distributions (features, pairs of features, ...) in NIST 302 are the same as those in NGI
- **Challenge**: *Access to NGI!*
- If special people are allowed NGI access: Find frequency of features; frequency of *pairs* of features within x mm; etc.
- Do these frequency distributions match?
- If not, what sample of NIST 302 prints is “representative” of NGI?

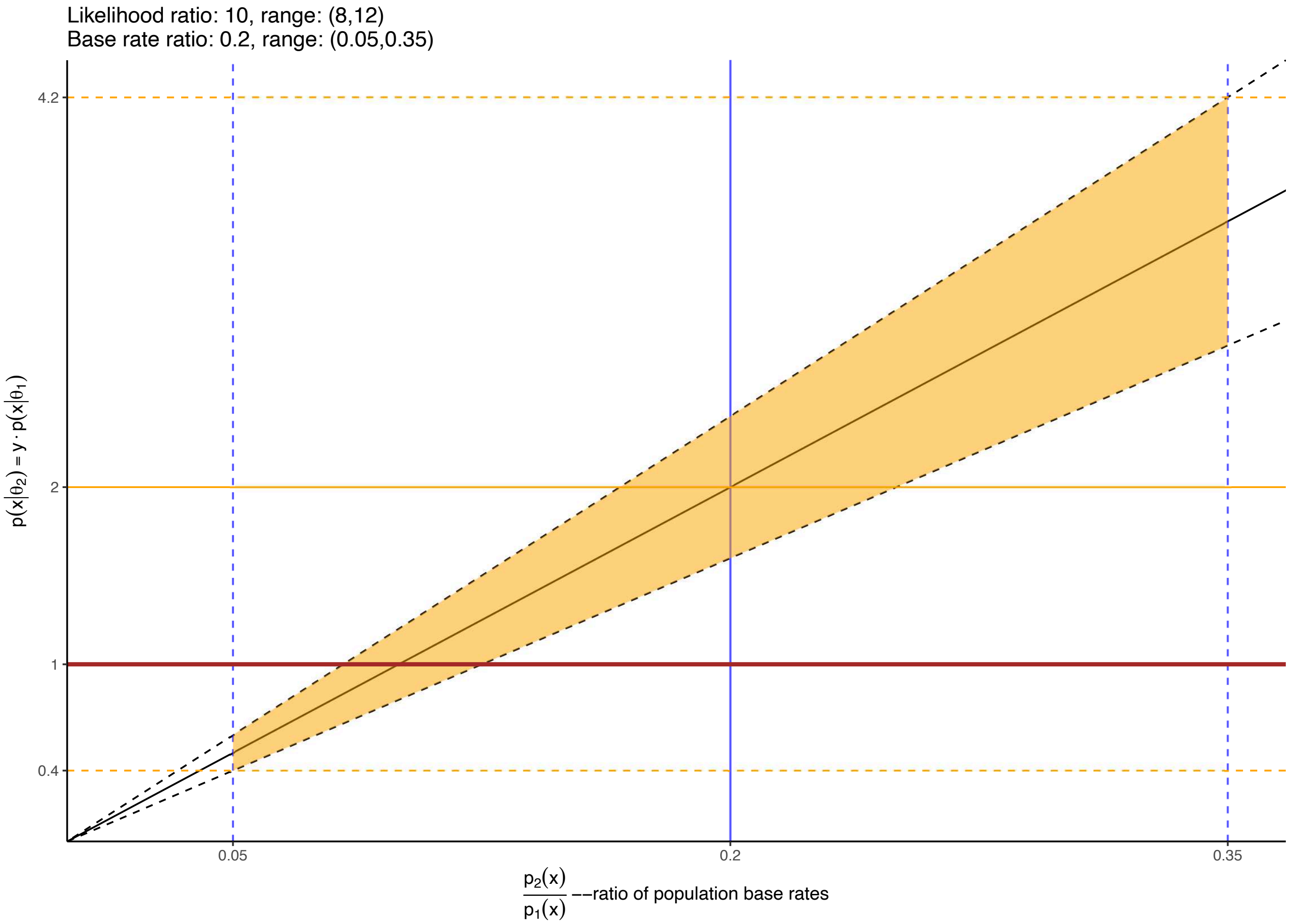
Same problem when developing algorithms for tumor detection in an image: When is a ‘sample’ of images “sufficiently representative” of the population of images?

Communicating results via informative displays

- KK's naïve ST/SA "classifier": **"LR"** = 5.8
- *What does "Likelihood Ratio" mean to a judge (much less jury)?*
- Even when *we* explain it correctly, frequently "LR" is *misunderstood* as "posterior odds." That's very bad.
- Can we provide a display that illustrates: *LR* and prior odds (both have *uncertainty*) and their consequent *posterior odds*?

Jordan Rodu: LR=10 (±2), Prior Odds 0.8 (±0.2) or 0.2 (±0.15)


Likelihood ratio: 10, range: (8,12)
Base rate ratio: 0.8, range: (0.6,1)
12
8
4.8
1
$p(x|\theta_2) = y \cdot p(x|\theta_1)$
0.6
0.8
1
$\frac{p_2(x)}{p_1(x)}$ --ratio of population base rates


Likelihood ratio: 10, range: (8,12)
Base rate ratio: 0.2, range: (0.05,0.35)
$p(x|\theta_2) = y \cdot p(x|\theta_1)$
4.2
2
1
0.4
0.05
0.2
0.35
$\frac{p_2(x)}{p_1(x)}$ --ratio of population base rates

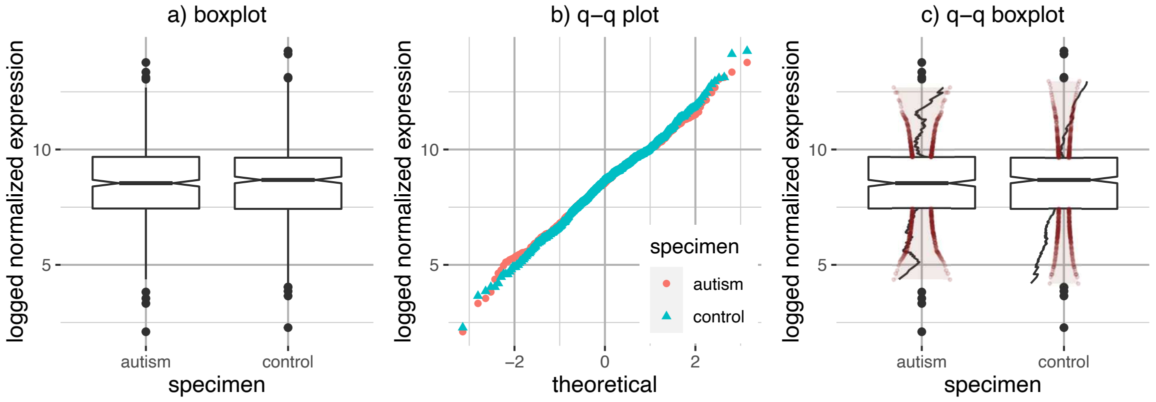
3. Why Efficient Statistical Computing?

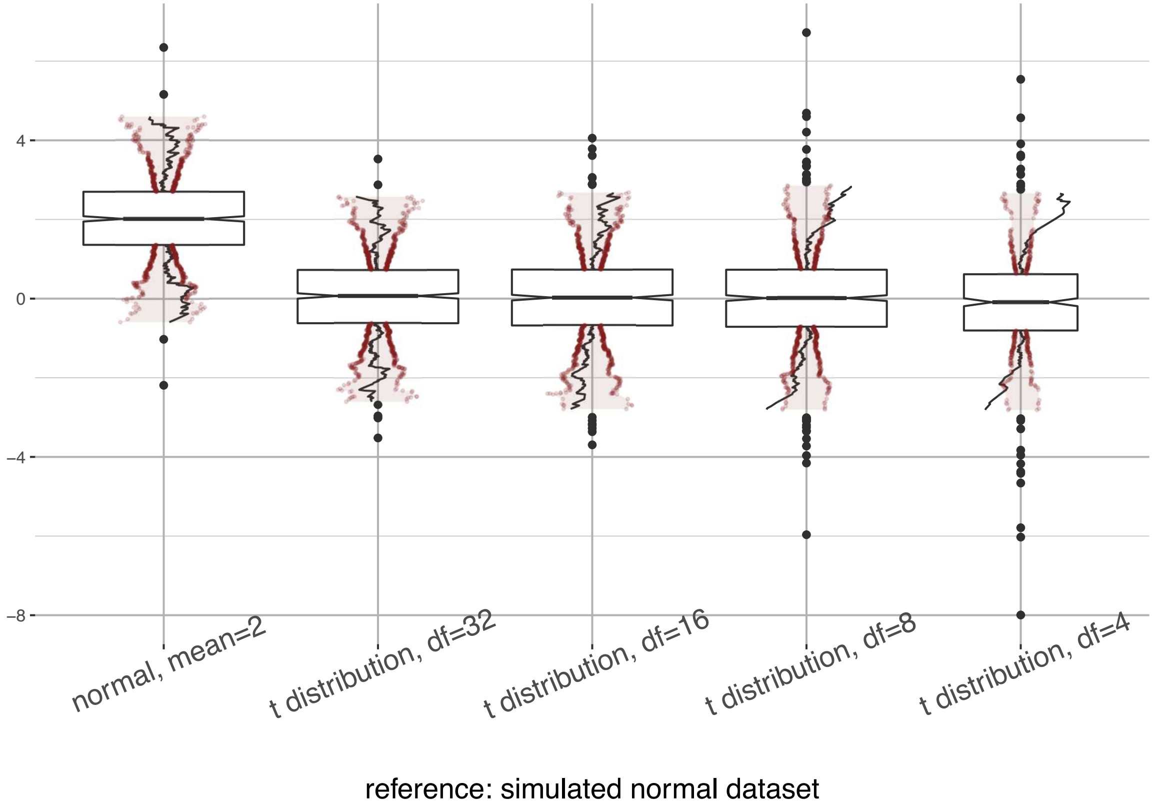
- Efficient stratified sampling from “big” data
- Estimate quantiles in streaming multivariate data
- Immediacy of results: medicine, credit card fraud detection, culprits, ... wherever time is critical
- Informative displays: Combining boxplots & qqplots in a single display for comparing multiple batches (Rodu & KK)
- “To ‘Learn’ from data for predicting future”

Prediction using “black-box” computational algorithms?

Rodu & KK: QQ boxplots (*JCGS* 2022)

- Multiple batches of data (years, conditions, regions)
- Boxplots: level and rough shape information
- QQ plots: tail behavior relative to a reference (usually Gaussian but possibly “historical” reference dataset)
- Combine into one plot
- Whiskers ⇒ Deviations between data & theoretical quantiles
- Confidence bands are easy for theoretical quantiles
- For reference dataset, need bootstrap
- Bootstrap will be computationally expensive for very large datasets; need efficient algorithms or subsampling schemes


a) boxplot
logged normalized expression
10
5
autism
control
specimen
b) q–q plot
logged normalized expression
10
5
−2
0
2
theoretical
specimen
autism
control
c) q–q boxplot
logged normalized expression
10
5
autism
control
specimen


4
0
−4
−8
normal, mean=2
t distribution, df=32
t distribution, df=16
t distribution, df=8
t distribution, df=4
reference: simulated normal dataset

4. To use or Not to use Black-box algorithms?

Jordan Rodu & Michael Baiocchi, "When black box algorithms are (not) appropriate" *Observational Studies* 9(2):79-101 (`doi.org/10.1353/obs.2023.0018`)

- Promises of AI: Accurate, fast predictions
  Many AI/ML algorithms use regression, Bayesian methods, ... with strong beliefs that they *will* predict well
- Dangers: Too much faith in accuracy, despite often trained on biased datasets; e.g. MIT Technology Review, 21-Jan-2019:
  "**AI is sending people to jail – and getting it wrong**
  Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past"

- “Black-box”: No real sampling statistical methodology
- Our usual statistical framework (model, parameter estimation, inference, CIs for estimates & predictions) do not apply
- Rodu & Baiocchi:
  When *are black-box algorithms appropriate?*

Focus on aspects of the **problem** (e.g., consequences of correct [incorrect] prediction) to which a solution is needed, *not* on the **algorithm** (e.g., functional form, assumptions, etc.)

Why use AI/black-box algorithms?

- Typical use: **classification**

  Healthy/Unhealthy; Low/Medium/High Credit (Flight) Risk; ...

- Statisticians tend to focus on *approaches*:

  How can we formulate the *algorithm* as a statistical problem?

- Friedman, Hastie, Tibshirani (2020): Additive logistic regression:

  A statistical view of boosting, *Ann Statist* 28(2): 337-407

Rodu and Biaocchi (*Observational Studies*, 2023)

- Statistical formulation for “black-box” algorithms may not be possible
- But we *can* provide a framework for the *problem*
- *Common Task Framework* has come to be the “arbiter” of “best” algorithms via hold-out data; cf. particle physics expts
- CTF appropriate for *outcome-* (vs *model-*) based reasoning

When is outcome-based reasoning appropriate?

Model reasoning (ex: logistic regression):

- focus on model covariates and assumptions
- ensure assumptions are legitimate, estimate model parameters
- Estimates based on existing data
- How will $\hat{f}$ perform on *future* data?
- Relate changes in model covariates (e.g., tall vs short) to changes & variations in outcomes
- Consideration of interaction between properties of *data* and properties of *algorithm*
- Comparison of algorithms: Estimation of parameters (bias, variance), predictions (accuracy, uncertainty)

Outcome reasoning: "*a new way to reason using data … that does not require slow-moving mathematical proofs*"

- Trained on existing data
- Performance evaluated on *hold-out* data
- Focus on correct outcome; i.e., correct prediction (classification)
- "Add" future data to current data to "tune" the algorithm
- Need neither a functional specification nor an *understanding* about the data

Common Task Framework is well-suited for this sort of development and validation of the algorithm

Rodu & Baiocchi: **Four conditions** that a **problem** must satisfy:

1. **M**easurement: A function of individual predictions and actual outcomes on *future* data (e.g., absolute or relative differences) can be measured

2. **A**daptibility: The algorithm can be adapted to the data on a *useful* timescale

3. **R**esilience: The *problem* can tolerate, or is resilient to, *accumulated error in predictions*

4. **A**gnosis: Outcome of algorithm can tolerate, or is agnostic to, potential discrepancies in stakeholders' prior beliefs

If the **problem** falls within MARA framework, “outcome reasoning” & “black-box” algorithms may be appropriate:

1. Measure discrepancy between algorithm’s output on future input) and (actual result): Monitor the error function

   Health care may *not* be suitable: No outcome following Rx

2. Adapt (update) the algorithm quickly to changes in input

   US Elections *not* suitable: electorate, priorities change each year

3. Resilience: Problem can tolerate errors

   Recommender systems: Bad recommendation not harmful;

   AI for sentencing or predicting recidivism/criminal risk *not* suitable

4. Agnosis: All stakeholders are comfortable with not knowing if algorithm did/didn’t incorporate certain data features

   SA/ST: does algorithm include “presence of petechaie”?

Rodu & Baiocchi argue that, if the **problem** does not meet all four MARA conditions:

Measurement, Adaptability, Resilience, Agnosis

then *Outcome Reasoning is not a good idea*:

- It is bound to fail when it encounters data that the algorithm has never seen before —
- even if trained to predict such data well as soon as it sees them
- Subpopulations or exotic data will continue to arise
- So "black-box" algorithms will continue to fail

Model reasoning is a better (safer, more quantifiable) choice on such problems - the majority of problems we face.

5. Summary: Statistics Now & Later

Some critical “Statistical Milestones”

- Stigler’s *The Seven Pillars of Statistical Wisdom*: summaries, $1/\sqrt{n}$, likelihood, comparison (via internal variation), regression, design, residuals
- Hypothesis testing framework (Neyman-Pearson 1933): Null/Alternative hypotheses
- Robust methods (Hampel 1974): Estimator, T, as a functional applied to an assumed distribution function, F_0; metric space on Fs, assess “robustness” via $T(F')$ for F' “near” $F_0 \Rightarrow$ framework for evaluating robust estimators

- Bootstrap framework (Efron 1979) for estimation & inference based only on *empirical* $\hat{F}$, not on *assumed* F
- False Discovery Rate (FDR; Benjamini & Hochberg 1995): Multiple testing framework by changing criterion, from "control probability of at most 1 false rejection" to "control *fraction* of false declared rejections"
- (Future? Candes' Knockoffs?)

They are so commonplace today that we forget how revolutionary they were for statistical practice at the time they were proposed.

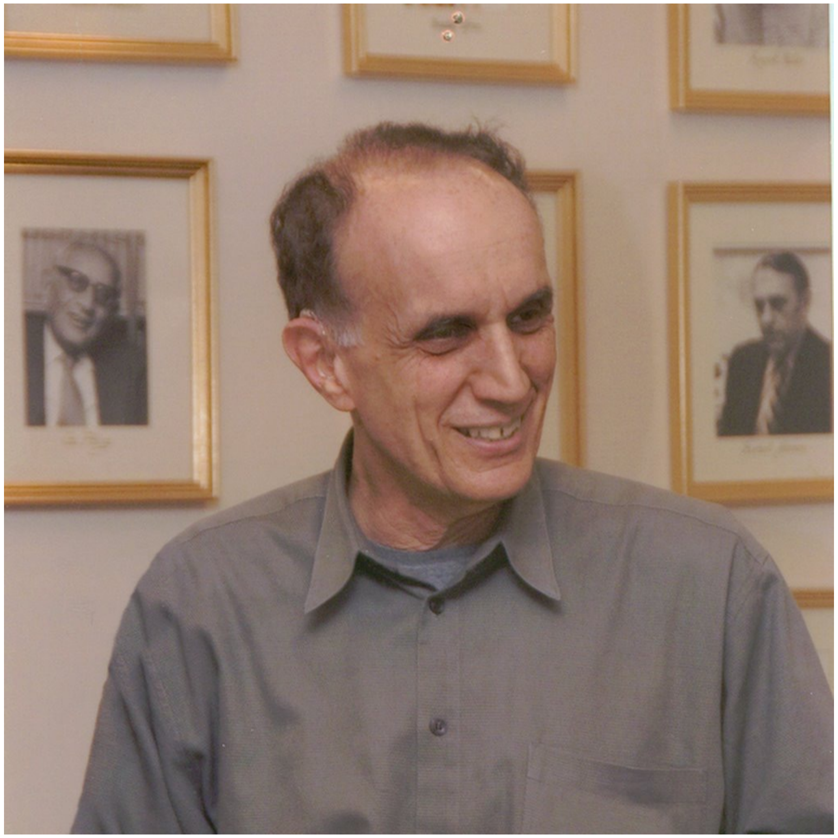
Do we need a new framework for “big-data” analysis?

Meng 2018 *AOAS*:

“*[I]t appears that the more we lament how our nutritious recipes are increasingly being ignored, the more fast food is being produced, consumed and even celebrated as the cuisine of a coming age. Indeed, some of our most seasoned chefs are working tirelessly to preserve our time-honored culinary skills, while others are preparing themselves for the game of speed cooking.*”

Can we reduce this blind devotion to “big data”?

Pandemic enhanced the reputation of Statistics discipline



2019 International Prize in Statistics Laureate Bradley Efron

"People have been predicting the demise of statistics for years.
First it was Computer Science. Then Operations Research. Then AI.
Then Expert Systems. Then Systems Engineering. Guess what - we're still here."

Statistical concepts needed for proper inferences in 'big data'

- Repetitious data vs Informative knowledge:
  "*We are drowning in information and starving for knowledge.* —
  R.D. Roger, from Hastie, Tibshirani, Friedman (2008), p.xi
- Sampling strategies
- Efficient computing: what to compute & display
- Informative displays of massive, streaming, data
- Robust methods to help flag "exotic" data segments
- Quantify steps of MARA framework in statistical terms:
  - Measurement: measure $d(y_{pred}, y_{obs})$
  - Adaptibility: tolerate large $d(Dataset_A, Dataset_B)$?
  - Resilience: Define "consequence" of error?
  - Agnosis: Measure importance of variable inclusion/exclusion?

AI/ML, “Data Science” still need Statistics

Meng 2018:

“*Fast food will always exist because of the demand … this is the very reason that we need more people to work on understanding and warning about the ingredients that make fast food (methods) harmful; to study how to reduce the harm without unduly affecting their appeal; and to supply healthier and tastier meals (more principled and efficient methods) that are affordable (applicable) by the general public (users).*”

We have work to do.

Thank you!

Some References

Donoho D (2017), 50 years of data science, *JCGS* 26(4):756-766

Efron B (2020), Prediction, Estimation, and Attribution, *JASA* 115(530):636-655

Meng X-L (2018), Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 U.S. Presidential election, *AOAS* 12(2),685-726

Rodu J, Baiocchi M (2023), When black box algorithms are (not) appropriate, *Observational Studies* 9(2):79-101 (`doi.org/10.1353/obs.2023.0018`)

Rodu J, Kafadar K (2022), The qq boxplot, *JCGS* 31(1):26-39 (`doi.org/10.1080/10618600.2021.1938586`)

Stigler SM (2016), *The Seven Pillars of Statistical Wisdom*, Harvard Univ Press.